Dynamics of Human Cooperation in Economic Games

Martin Spanknebel [*1] and Klaus Pawelzik [†1,2]

[1]Institute for Theoretical Physics, University of Bremen
[2]Center for Cognitive Sciences, University of Bremen

Human decision behaviour is quite diverse. In many games humans on average do not achieve maximal payoff and the behaviour of individual players remains inhomogeneous even after playing many rounds. For instance, in repeated prisoner dilemma games humans do not always optimize their mean reward and frequently exhibit broad distributions of cooperativity. The reasons for these failures of maximization are not known. Here we show that the dynamics resulting from the tendency to shift choice probabilities towards previously rewarding choices in closed loop interaction with the strategy of the opponent can not only explain systematic deviations from 'rationality', but also reproduce the diversity of choice behaviours. As a representative example we investigate the dynamics of choice probabilities in prisoner dilemma games with opponents using strategies with different degrees of extortion and generosity. We find that already a simple model for human learning can account for a surprisingly wide range of human decision behaviours. It reproduces suppression of cooperation against extortionists and increasing cooperation when playing with generous opponents, explains the broad distributions of individual choices in ensembles of players, and predicts the evolution of individual subjects' cooperation rates over the course of the games. We conclude that important aspects of human decision behaviours are rooted in elementary learning mechanisms realised in the brain.

[*]martin.spanknebel@uni-bremen.de
[†]pawelzik@neuro.uni-bremen.de

Humans and animals regularly fail to achieve maximal payoff even in simple choice situations and also when allowed for trying them out with many moves. Such observations have triggered a plethora of research into the lack of 'rationality' in human and animal decision behaviour. Deviations from utility maximization were explained by psychological peculiarities including e.g. probability matching [1] and mis-estimations of utilities [2]. Also, auxiliary motives or emotions can have a substantial influence on decision behaviour which cannot always be excluded experimentally [3]. Last, not least exploration and learning mechanisms are increasingly suspected to contribute substaintially to deviations from strict rationality in repeated games (for an encompassing review see [4]). Here, wide, skewed, and U-shaped distributions of choice propensities were occasionally observed [1, 4, 5, 6].While these findings hint to the existence of multiple attractors for the choice propensities in analogy to similar considerations in evolutionary game theory [7, 8], this link has not been established yet for human learning in economic games.

Social dilemma games have served as a paradigm in theoretical research of the evolution of cooperativity. Being particularly well understood they provide a testbed for investigating human deviations from optimal decision behaviour. In the prisoner's dilemma two players, named X and Y, are playing a game in which they can either cooperate or defect. In the case of mutual cooperation both get a reward r_{CC}. If X defects and Y cooperates, X gets an even larger reward r_{DC} while Y gets an lower reward r_{CD}. If both players defect, both get an reward r_{DD}. These rewards must satisfy two equations: $r_{DC} > r_{CC} > r_{DD} > r_{CD}$ leads to a Nash Equilibrium of mutual defection, while $2r_{CC} > r_{DC} + r_{CD}$ guarantees, that mutual cooperation has the best global outcome. The latter condition establishes the dilemma between individual gain for defection and larger common benefit for cooperation. In the following we use the standard values defined in table 1, which Axelrod used for his famous computer tournament[9].

Table 1: Definition of the reward r

		Computer: X	
		cooperate $x = C$	defect $x = D$
Human:Y	cooperate $y = C$	0.3 (r_{CC}) / 0.3 (r_{CC})	0.5 (r_{DC}) / 0.0 (r_{CD})
	defect $y = D$	0.0 (r_{CD}) / 0.5 (r_{DC})	0.1 (r_{DD}) / 0.1 (r_{DD})

Recently a new class of strategies for the prisoners dilemma was introduced by Press and Dyson [10], the so called Zero Determinant (ZD) Strategies. ZD strategies are memory-one strategies, i.e. they only use the information from the last round to base their decision. It can be shown that against ZD strategies longer memory strategies cannot have any advantage over memory-one strategies. ZD strategies have two further characteristics: (i) they enforce a linear relationship between the rewards of the players; (ii) for positive slope of this relation the best strategy for the co-player, meaning the
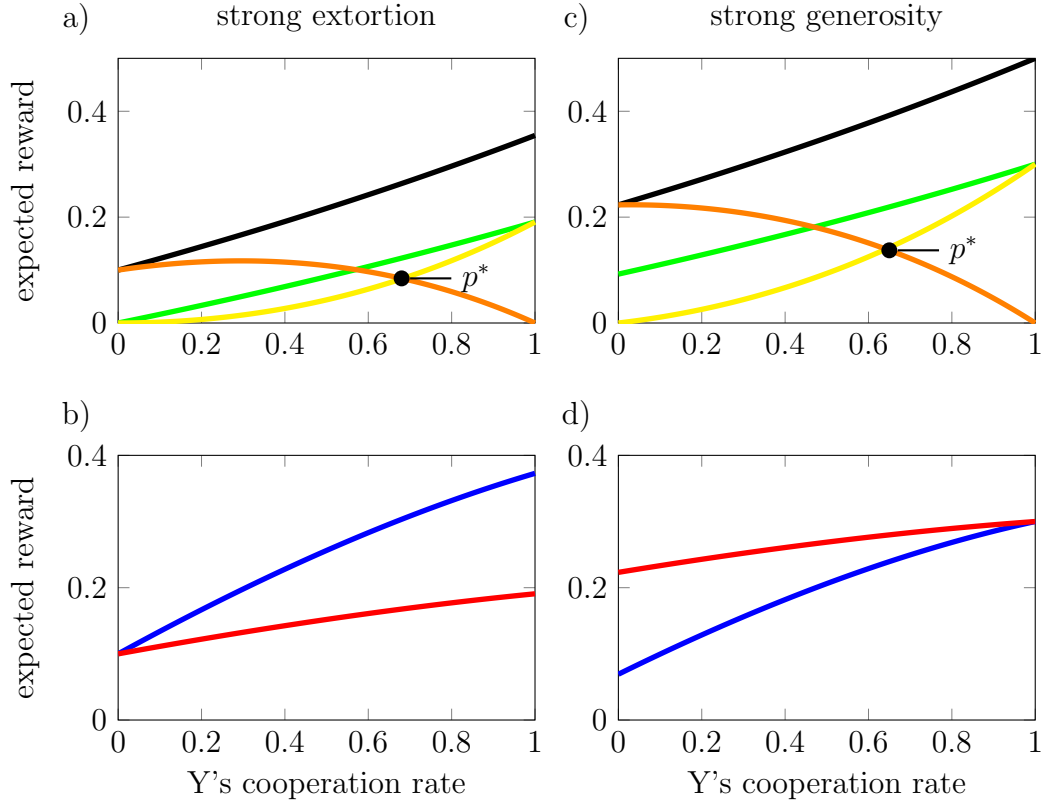
Figure 1: Panel a) and c) show conditional expected rewards for strong extortion and strong generosity, respectively, for the context free player using a single cooperation rate p. The black line is the expected reward r(p—y = C) given the player cooperated and the green line the expected reward r(p—y = D) given the player defected. The yellow and orange lines are these conditional expected rewards weighted with their corresponding probabilities p and 1-p yielding the expected rewards from cooperative and defective acts, respectively. The learning rule has an unstable fixed point p^* where these functions cross. Panel b) and d) show the total expected rewards of the player (red) and of the respective opponent ZD strategy (blue).

strategy with the highest reward is total cooperation. In this paper we will concentrate on two extreme types of these ZD strategies: An extortion strategy which dominates any adaptive opponent [10]and a generous strategy which ensures that her own reward does not exceed the reward of the co-player [11]. Figure 1 show the expected reward of a purely stochastic player (Y) having a single fixed cooperation probability playing when against the extortion 1b) or the generous strategy 1d) of X. For this simple strategy of Y, that takes no memory into account, but also for all other possible strategies of Y, total cooperation always leads to the maximum reward [10].

An elegant experiment by Hilbe et al. demonstrated a particularly striking lack of cooperation in humans [12] when repeatedly playing prisoner dilemma games against zero-determinant strategies that extort their opponents [10]. In contrast, when playing against generous ZD-strategies, human cooperation rates increased with time. As an explanation for this substantially sub-optimal behaviour against extortionists a desire for punishment was offered.

This experiment also revealed the large diversity of individual choice propensities ('cooperativities') sometimes leading to bimodal ('U-shaped') distributions after a series of moves, where some participants ended up always defecting while others always cooperated [12, supplement, Fig. 1].

Obviously, the dynamics of choice propensities during the course of a game can lead to opposite behaviours in different individuals. We wondered if the coevolution of learning in the human player in interaction with the opponent's strategy could yield a dynamics that would explain this wide spread of individual behaviours.

Adaptations of human and animal behaviour in repeated games have been studied extensively in economics [13, 14, 15]. Since the pioneering work of Pavlov more than hundred years ago research into reward dependent learning of choice preferences brought about a range of hypothetical learning rules based on success-dependent reinforcement of behaviour. It turned out that adaptations of animal as well as human behaviours are far from trivial even in very simple choice situations (for games see the review [4]). Already the phenomenology of Classical Conditioning can be explained only partially by the famous Rescorla-Wagner learning rule [16]. Therefore, more sophisticated algorithms for reinforcement learning were introduced, among which TD-learning [17] is now a leading paradigm in behavioural brain research and the Bush-Mosteller-Rule (BMR) [18] is frequently used in Psychology.

All psychological learning rules share the basic principle of reinforcing choice probabilities for successful choices and/or weakening choice probabilities for negatively rewarded decisions (Thorndike's famous 'law of effect',[19]). Systematic biases either inherent in the particular learning rule or in specific parameter settings were frequently used to explain systematic deviations from optimal behaviour. As a representative example we here only cite a recent careful study in which success dependent learning effects in professional basketball players were explained by different learning rates for updating the respective choice probabilities [20].

Several studies investigated adaptations of choice probabilities in their interaction with

4

the dynamics of the environment. An important example here is the work of Herrnstein [1], which shows that neither probability matching nor payoff maximization can explain human decision behaviour when the response of the environment is history dependent. To our knowledge the first mathematical analysis of the dynamics of choice probabilities was performed by Izquierdo and co-workers [21], who used the Bush-Mosteller-Rule to investigate the putative flow of cooperation probabilities of two players in social dilemma games. While their formal method is interesting from a theoretical perspective, the concrete predictions of this study are difficult to test since this would require knowledge of the strategies used by both human players which is not available.

The case considered by Hilbe et al. [12] is more useful in this respect since the opponent's fixed zero determinant strategies are analytically fully transparent [10]. We use the data from these experiments as a benchmark for investigating the evolution of individual choice probabilities in interaction with a dynamic environment.

Results

We model the choice behaviour $y(t) \in \{C, D\}$ of a subject Y in prisoner dilemma games using a stochastic strategy based on cooperation probabilities $p_Q(t)$. Given the context Q, and the decision at (discrete) time $t = 1, ..., T$ the player receives a reward $r(t)$ which depends on the behaviour $x(t) \in \{C, D\}$ of the opponent X. In general Q can consist of any information available during the game.

Here we restrict our analysis to three cases: Q representing a) no context, b) the choice $x(t-1)$ of the opponent X in the last round, and c) the combination $(x(t-1), y(t-1))$ of the choices of X and Y in the last round. Since taking more history into account cannot help Y to increase her income when playing against Zero-Determinant (ZD) strategies [10] it is not considered.

The choice probabilities $p_Q(t)$ are adapted according to the success of behaviour by a simple learning rule :

$$
p_Q(t+1) = p_Q(t) + \begin{cases} +\varepsilon_C r(t)\delta_{Q,Q'} \text{ for } y = C \\ -\varepsilon_D r(t)\delta_{Q,Q'} \text{ for } y = D \end{cases} \tag{1}
$$

where $\varepsilon_{C,D}$ are the learning rates in the case of cooperation and defection, respectively. $\varepsilon_C > \varepsilon_D$ leads to a bias b towards cooperation. $\delta_{Q,Q'}$ is the Kronecker-Delta which equals one only if Q matches the context Q' that was present at time t, and is zero otherwise.

With no context (case a) the index Q can be dropped. The expected change $\langle \Delta p \rangle = \langle p(t+1) - p(t) \rangle$ of the cooperation probability p can be calculated (see Methods) and depends on the difference of the expected reward for cooperation $\bar{r}_C$ and that of defection $\bar{r}_D$. Fig. 1, a,c shows the results for the games against extortionist and generous
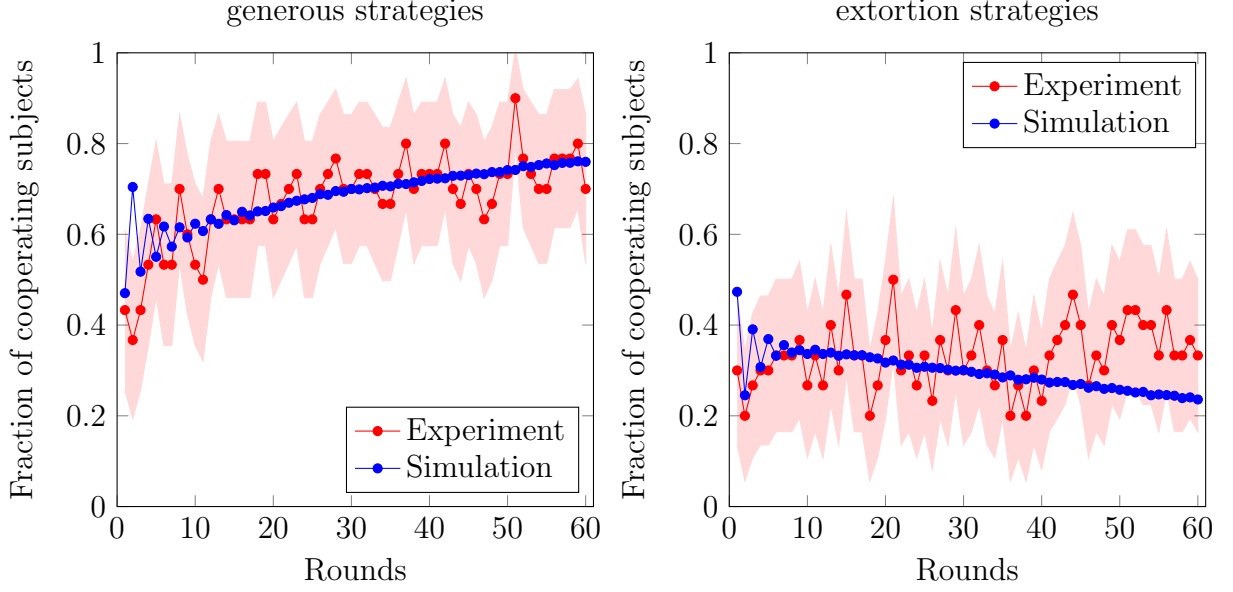
Figure 2: Fraction of cooperating subjects during the course of the game from the experiments by Hilbe et al. [12] (red) and in simulations (blue) with $N = 10201$ players. The pink shaded region represents the standard errors from the experiments.

opponents, respectively. The fixed points $p^* : r(\bar{p}^*)_C = r(\bar{p}^*)_D$ for both cases are unstable, players with initial p above this point will tend to increase their cooperation probabilities and vice versa. In ensembles of players with widely sprad initial conditions the distribution will split, the respective cooperation probabilities will move towards the boundaries, and after many rounds one expects 'U-shaped' distributions of cooperation propensities. Ensembles with initial conditions above the fixed point will converge towards cooperation, which has indeed been observed [22]. However, since the fixed points for extortion and generosity are quite similar this case is not able to explain the results from Hilbe et al. [12].

Case a) serves to demonstrates the basic effect: reward dependent adaptation of p can induce multistability and not neccessarily leads to maximization For this case of no history dependency of the players cooperation propensity the difference between extortionist and generous opponents is small. Since the unstable fixed point are nearly at the same position($r(p^*) = 0.69$ in the extortion and $r(p^*) = 0.65$ in the generous case) human subjects would act similar if playing against the different treatments.

Case b) with $Q \in \{C, D\}$ is more useful to explain the results by Hilbe et al. even if no cooperation bias is present ($\varepsilon_C = \varepsilon_D$). Case c) leads to a similar dynamic as case b) and is therefore not shown. Figure 2 compares the fraction of cooperating subjects during the course of the game playing against extortion and generous treatments with the respective simulations of the model. In the experiment [12] the humans start with a cooperation rate of 30 - 40 %. In the generous treatment the subjects behave "rationally", i.e. they optimise their strategy towards cooperation. At the end they
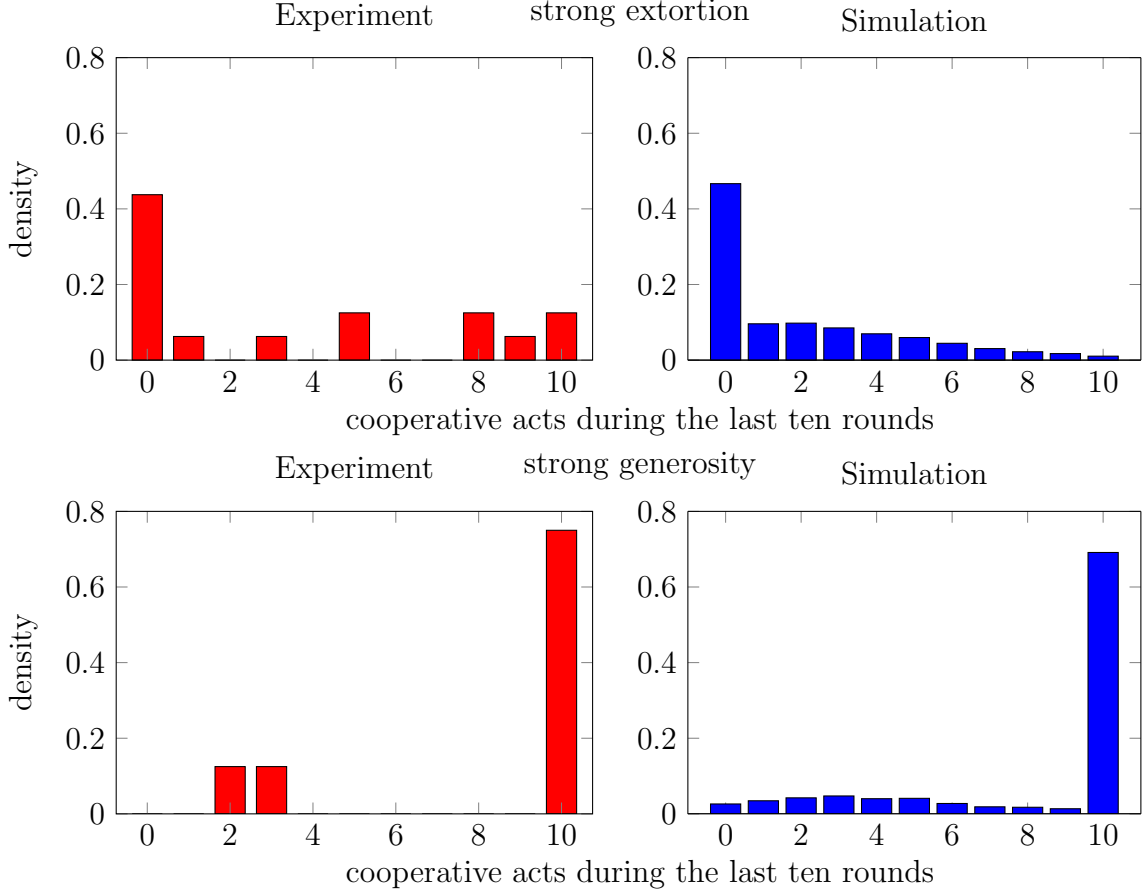
6

Figure 3: Cooperation acts during the last ten rounds. Left part of the figure taken from [12].

reach a cooperation probability of 70 - 80 %, which leads to a mean payoff of 0.27 €
which is not far away from the best possible payoff 0.30 €. Against extortion strategies
humans behaved differently. Their cooperation probability dropped to low values (30 -
40 %) and they received only a mean reward of 0.15 € which is quite far away from the
maximally possible payoff 0.22 €. The simulations lead to a very similar behaviour, in
the generous case the cooperation probability starts at 47 % and goes up to 76 %. In
the extortion case the agents also start with a cooperation probability of 47 % which
decreases during the course of the game and ends with a probability of 24 %. In both
cases the respective propensities of cooperation are fully consistent with the data, with an
increasing cooperation probability in the generous and a slightly decreasing cooperation
probability in the extortion case.

Figure 3 compares the behaviour of the model after 60 rounds with the results from
Hilbe et al [12]. In the strong extortion case the results of the model match the experi-
mental data, the peak at zero has nearly the same height. The generous cased matches
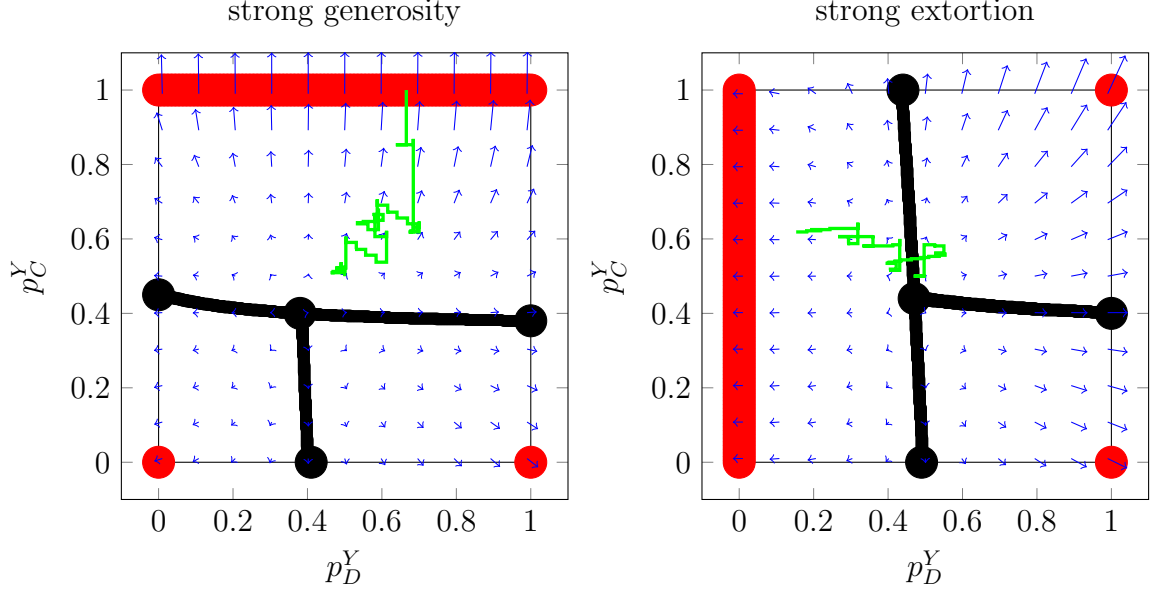qualitatively, the large peak at ten appears in both, the simulation and the experiment,

Figure 4: Flow in the strong extortion and generous case with the unstable (black dots) and stable manifolds(red dots and lines) and the corresponding zones of attraction(black lines). The green curves are exemplary movements of single agents with the starting point $(0.5, 0.5)$ over 200 rounds

but in the simulation this peak is smaller than in the experiment.

To understand the dynamics of the model we calculate the expected changes of both choice probabilities p_C and p_D (see Methods) and display them depending on p_C and p_D as a vector field (i.e. the flow in choice probability space as in Izquierdo et al. [21]). Figure 4 shows the unstable and stable fixed points and the corresponding basins of attraction of this flow. There are four unstable fixed points and three stable manifolds with basins of attraction that are delimited by the nullclines.

For the strongly generous opponent we find a single unstable fixed point within the preference space at $(p_D^Y, p_C^Y) = (0.38, 0.4)$. It determines an edge of the basin of attraction for the stable fixed point at the lower left edge at $(0,0)$, together with the fixed point at the left boundary at $(0, 0.45)$ and at the lower boundary at $(0.41, 0)$. This latter stable fixed point is the state of permanent defection. The basin of attraction for the stable fixed point at the lower right edge at $(1, 0)$ is formed by the central and the left fixed point and the fixed point at the right boundary at $(1, 0.38)$. At this fixed point the agents always take the different action as the opponent did in the last round, so we can call this behaviour "anti-Tit-for-Tat".

The stable manifold most relevant for the experiment with generous opponents is the upper one at $p_C^Y = 1$. The corresponding basin of attraction is bounded by the left, central and right unstable fixed point. On the one hand the basin of attraction of this manifold has the largest overlap with the initial conditions (see Methods), so most of the

Table 2: Cooperation probabilities of the ZD strategies. p_{CD}^X is the cooperation probability of the ZD strategy if the ZD strategy has cooperated and the opponent has defected in the previous round.

Strategies	p_{CC}^X	p_{CD}^X	p_{DC}^X	p_{DD}^X
strong extortion	0.692	0.000	0.538	0.000
strong generous	1.000	0.182	1.000	0.364

agents will go up towards this stable manifold. At the other hand agents in this manifold always cooperate if the opponent has cooperated in the last round, but because generous strategies also behave like this (see table 2) this is a point of permanent mutual cooperation. This explains why nearly every human in this experiment and the largest fraction of agents in the simulation cooperated permanently at the end of the game (see figure 3).

For the strong extortion opponents the fixed points are similar to those of the strong generous case. However, the central unstable fixed point (here at $(0.47, 0.44)$) together with the upper $(0.44, 1)$, the lower $(0.49, 0)$ and the right $(1, 0.4)$ unstable fixed points now divide the space into three distinct basins of attraction. All agents, which are initialized in the upper right basin will tend to go to the stable fixed point at $(1, 1)$ of permanent cooperation. In contrast, the agents which start in the lower right basin will move towards the anti-Tit-for-Tat fixed point at $(1, 0)$.

The agents, who start at the left basin will tend to go to the large stable manifold at $p_D^Y = 0$, which is most important for understanding the experimental results for this case. On the one hand the basin encompasses the initialization area and on the other hand the agents always defect if the opponent has defected in the previous round. Since extortion strategies do the same (see table 2) this leads to a state of permanent mutual defection. This explains why the largest fraction of subjects in the experiment and agents in the simulation prefer defection at the end of the game(see figure 3).

A critical prediction of our model is that the evolution of the players choices should depend on their respective individual initial conditions. We tested this property of our model by comparing the final cooperation rates of individual subjects with the corresponding predictions of our model which were based on the subjects' observed choices during the first ten rounds (see Methods). Figure 5 summarises the results for all games and every subject. The linear regression (solid line) leads to a function with a slope of 0.69 and a y-intercept of 0.26.

In principle, the flow field can be estimated directly from experimental data which will allow to experimentally test different hypothetical learning rules. To demonstrate this, we used a naive method (see Methods) with computer generated data and found that a good match with the analytic flow can be achieved already from 500 agents in a game of 200 rounds (figure 6).
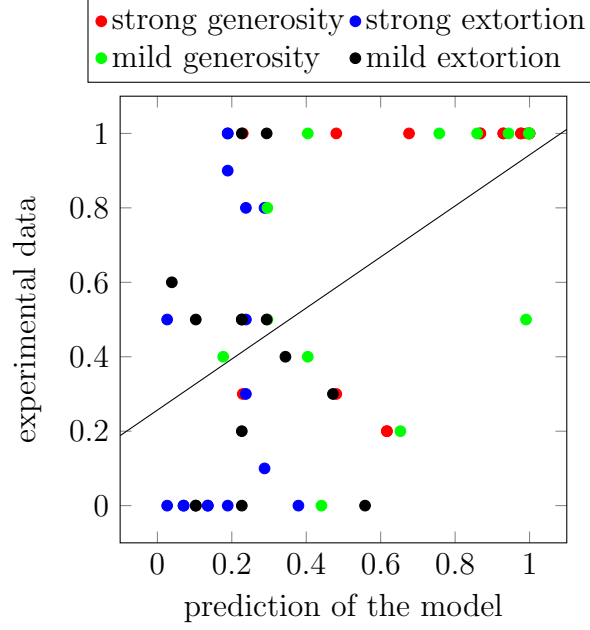
9

Figure 5: Fraction of cooperative acts during the last ten rounds of each subject in all 4 ZD-treatments from the experiments of Hilbe et al. [12] versus the predictions of the model. The solid line is the linear regression having a slope of 0.69 and a y-intercept of 0.26.

Discussion

The dynamics of choice probabilities was analysed for a simple model of learning in subjects playing iterated prisoner dilemma games against zero-determinant strategies. The flow in choice preference space exhibits characteristic basins of attraction: depending on their initial readiness to cooperate after many rounds of the game players tend to either always cooperate or to always defect. Thereby small individual differences in cooperation rates may lead to opposite behaviours in the long run. While the insight that learning can lead to extreme behaviours is not new [21, 23], the general method introduced here allows to determine the potential influence of the dynamics on systematic deviations from optimal behaviour for all environments that depend on few past moves. For instance, we suspect that many of the systematic deviations from both, utility maximization and matching observed experimentally as in [1] can be explained by our framework.

When playing against opponents with zero-determinant strategies a longer memory beyond the last move cannot improve the payoff [10], an intriguing fact, which makes these games a benchmark for investigating the joint dynamics of learning and environment. With a plausible initial distribution of cooperation probabilities the model is able to quantitatively explain the full phenomenology of the population data gathered by Hilbe et al. [24]. In particular, it reproduces the average dynamics of cooperation rates and explains the wide distributions of cooperativity among individual players.

Our results are robust with respect to many details of the learning rule. As a test we applied the full Bush-Mosteller Rule [18] and obtained similar results when the aspiration parameter in the BMR was small. The aspiration is an offset to which the received payoff is compared and only the difference is used for reinforcement of the choice probabilities. Since the subjects in the experiment of Hilbe et al. were payed according to the absolute payoff of the prisoners dilemma game and had no explicit costs setting this parameter to zero is a plausible choice. If, however, the subjects would need to compare the received rewards to some explicit costs our model would predict the emergence of stable fixed points in preference space located at intermediate values of cooperation rates. This prediction of our approach could easily be tested in future experiments.

A central prediction of our model that can be tested already on the data available from [24] is that the asymptotic cooperation rates of individual players should depend on their initial tendency to cooperate. This critical prediction was confirmed to a striking degree: based on the sequence of choices of individual subjects during the first ten rounds of the games the model's predictions closely matched the average fraction of cooperative acts in the last ten rounds (figure 5). Prediction improved with a non-zero bias for successful cooperation in the learning rule, which is equivalent to assuming an additional value of mutual cooperation [25]. These results demonstrate that the proposed learning mechanism not only qualitatively captures the dynamics of choices. We did not attempt to further optimize the two free parameters of the model (learning rate and cooperation bias), neither did we perform an extensive search for better initializations of the cooperation rates which obviously can only improve these results. Also, we did not consider different individual learning rates, which can further improve the match to experimental data, however, at the cost of additional model parameters.

We believe, however, that the prediction performance of our model will improve more substantially if the initial cooperation propensity and the learning rate could be estimated directly for each subject before the game starts. Then, given the model is correct, only the noise intrinsic to players and opponents would limit its predictability. In principle, also the flow itself can be estimated from experimental data. Using a naive method flow estimation in simulations (see Methods) became significant with 500 players making 200 moves. This second approach demonstrates that experiments are within reach, that would directly test hypothetical learning rules. We plan to perform these tests in the next future. Because the method of determining the flow from groups of players can be implemented for any learning rule it can in principle also be used for determining characteristics of learning implemented in subjects from selected groups. We speculate, that it would thereby become possible to investigate the relation of pathological behaviour (as e.g. in addiction) to systematic aberrations of learning.

Taken together, our results suggest a fundamental role of the joint dynamics of learning and environment for explaining experimentally observed deviations from optimal decision behaviour that has previously been overlooked.

Methods

Model

Since the learning rule (1) can lead to probabilities smaller than zero and larger then one we always set the cooperation probability to zero and one, respectively, if these boundaries are passed. The learning rates are $\varepsilon_C = 0.09375$ and $\varepsilon_D = 0.03125$ and all simulations were performed with $N = 101^2 = 10201$ agents. For the initial conditions we assume, that human beings rather cooperate if there opponent has cooperated as well respectively defect if the opponent has defected as well. That is we choose a uniform distribution with the borders $0 < p_D^Y < 0.45$ and $0.45 < p_C^Y < 1$. In the first round we choose one of the two cooperation probabilities randomly.

Mean Field Theory

To understand the results of the model, we calculate the expectation value $F_k = < \Delta p_k^Y >$ of equation (1). $\sigma_{Y,t}$ only depends on Y's behaviour in the current round, the reward r_{xy} depends on both players behaviour in the current round and the Kronecker delta $\delta_{x'k}$ depends on X's behaviour in the previous round. The behaviour of X and Y in the current round are called x and y, respectively, and in the previous round x' and y'. F_k only depends on x, y and x'. To calculate this expectation value we need the joint probability $\pi(x, y, x')$:

$$F_k \sim \sum_{\{x,y,x'\} \in \{C,D\}} \pi(x, y, x') \sigma_y r_{xy} \delta_{x'k}. \tag{2}$$

To calculate $\pi(x, y, x')$ we sum over the joint probabilities which include the full information of the last and current round:

$$\pi(x, y, x') = \sum_{y' \in \{C,D\}} \pi(x, y, x', y') = \sum_{y' \in \{C,D\}} \pi(x, y | x', y') S_{x'y'}. \tag{3}$$

$S_{x'y'}$ is the probability of the occurrence of the state (x', y'). A repeated prisoner's dilemma can be seen as a Markov process. The only problem is, that Markov processes assume constant transition probabilities which is not the case in our model because of the learning rule. We now assume that this process is a quasistatic process, which allows us to use the steady state probabilities of the Markov process. These state probabilities can easily be calculated by solving the following eigenvalue equation:

$$\mathbf{S} \begin{pmatrix} p_{CC}^X p_C^Y & p_{CD}^X p_C^Y & p_{DC}^X p_D^Y & p_{DD}^X p_D^Y \\ p_{CC}^X (1 - p_C^Y) & p_{CD}^X (1 - p_C^Y) & p_{DC}^X (1 - p_D^Y) & p_{DD}^X (1 - p_D^Y) \\ (1 - p_{CC}^X) p_C^Y & (1 - p_{CD}^X) p_C^Y & (1 - p_{DC}^X) p_D^Y & (1 - p_{DD}^X) p_D^Y \\ (1 - p_{CC}^X)(1 - p_C^Y) & (1 - p_{CD}^X)(1 - p_C^Y) & (1 - p_{DC}^X)(1 - p_D^Y) & (1 - p_{DD}^X)(1 - p_D^Y) \end{pmatrix} = \mathbf{S} \tag{4}$$

where p_{CD}^X is the cooperation probability of the ZD strategy, if the ZD strategy has cooperated and the human opponent has defected in the previous round. Quasistatic processes in general are processes which have an infinitely long break after each parameter change in which the system can equilibrate so that there is a permanent equilibrium. In our case of a repeated prisoner dilemma this means that after every round with a parameter change we assume an infinite number of rounds with no parameter change so that the system can equilibrate.

The probability $\pi(x, y|x', y')$ is the probability that in the current round the state (x, y) occurs, if in the previous round the state (x', y') occurred. This probability can be calculated with the known cooperation probabilities of X and Y:

$$\pi(C, C|x', y') = p_{x'y'}^X p_{x'}^Y \tag{5}$$
$$\pi(D, C|x', y') = (1 - p_{x'y'}^X) p_{x'}^Y \tag{6}$$
$$\pi(C, D|x', y') = p_{x'y'}^X (1 - p_{x'}^Y) \tag{7}$$
$$\pi(D, D|x', y') = (1 - p_{x'y'}^X)(1 - p_{x'}^Y) \tag{8}$$

Prediction of Final Cooperativity

For predicting final cooperation rates c_e^i from initial cooperation rates c_s^i we calculate numerically the probability $P(c_s, c_e)$ of the model that a randomly chosen agent has the cooperativity (c_s, c_e). With this probability we can calculate the conditional probability $P(c_e|c_s)$

$$P(c_e|c_s) = \frac{P(c_s, c_e)}{\int_0^1 P(\tilde{c}_s, c_e) d\,\tilde{c}_s} \tag{9}$$

and the conditional expectation value

$$E(c_e|c_s) = \int_0^1 P(\tilde{c}_e|c_s)\tilde{c}_e d\,\tilde{c}_e \tag{10}$$

Plotting the conditional expectation value of the predicted cooperation rates in the last ten rounds versus the experimental data $(E(c_e|c_s^i), c_e^i)$ should lead to a distribution similar to the angle bisector if the model is correct.

Measurement of the Flow

To check whether the learning rule is correct or not we want to estimate the flow from experimental data. Therefore we divide the sequence into several parts (which can overlap) and calculate the mean value for p_C^Y and p_D^Y in each part.

$$\bar{p}_C^Y = \frac{n_{CC'}}{n_{CC'} + n_{DC'}} \tag{11}$$
$$\bar{p}_D^Y = \frac{n_{CD'}}{n_{CD'} + n_{DD'}} \tag{12}$$

e.g. $n_{CD'}$ is the number of times the human subject cooperates in the current round, if the computer has defected in the previous round. If the sequences are long enough we can assume, that the differences of the mean values is caused by the change of the cooperation probabilities and not by noise. In the case that one agent only cooperates or defects it is only possible to estimate one of the two probabilities. In this case we assume that the other cooperation probability is unchanged and we use the probability of the previous part. If this appears at the beginning of the game we take the next probability.

We tested this very simple method with computer generated data of 500 agents in a game of 200 rounds. This is a large, but realistic experimental number of rounds. If the human subjects have e. g. 15 seconds for each decision, which is quite much for pressing one of two buttons, they would need 50 minutes. An experiment with 500 human subjects is a lot of work, but every subject plays independently against a computer. The humans do not have to play at the same time, so the amount of work increases linearly with the number of human subjects.

The results of this measurement are shown in figure 6. To determine the quality of this method we calculate the coherence of the analytic flow $\mathbf{F}_\mathrm{a}$ and the measured flow $\mathbf{F}_\mathrm{n}$:

$$c = \frac{\langle \mathbf{F}_\mathrm{a} \cdot \mathbf{F}_\mathrm{n} \rangle^2}{\langle \mathbf{F}_\mathrm{a} \cdot \mathbf{F}_\mathrm{a} \rangle \langle \mathbf{F}_\mathrm{n} \cdot \mathbf{F}_\mathrm{n} \rangle} \tag{13}$$

The largest coherences were found at a part length of 110 rounds and an overlap between these parts of 90 % (so part one goes from round 1 to 110, part two from round 12 to 122 etc.). The coherence for the shown examples are 0.40 and 0.53, which are typical for the strategies. Especially at the border the analytic and measured flow do not match well. This is cause by the boundary condition which is not used in the calculation of the flow but in the computer generated data.

The coherences of the extortion cases are significantly higher than in the generous cases. This is caused by the left part of the flow ($p_D^Y = 0 \dots 0.5$) where nearly all arrows are parallel, so errors in p_C^Y does not lead lo a lower coherence.

Acknoledgements

References

[1] Herrnstein RJ, Loewenstein GF, Prelec D, Vaughan W (1993) Utility Maximation and Melioration: Internalities in Individual Choice. *J Behav Decis Mak* 6:149–185.

[2] Kahneman D (2003) Maps of Bounded Rationality : Economicst Psychology for Behavioral. *Am Econ Rev* 93:1449–1475.
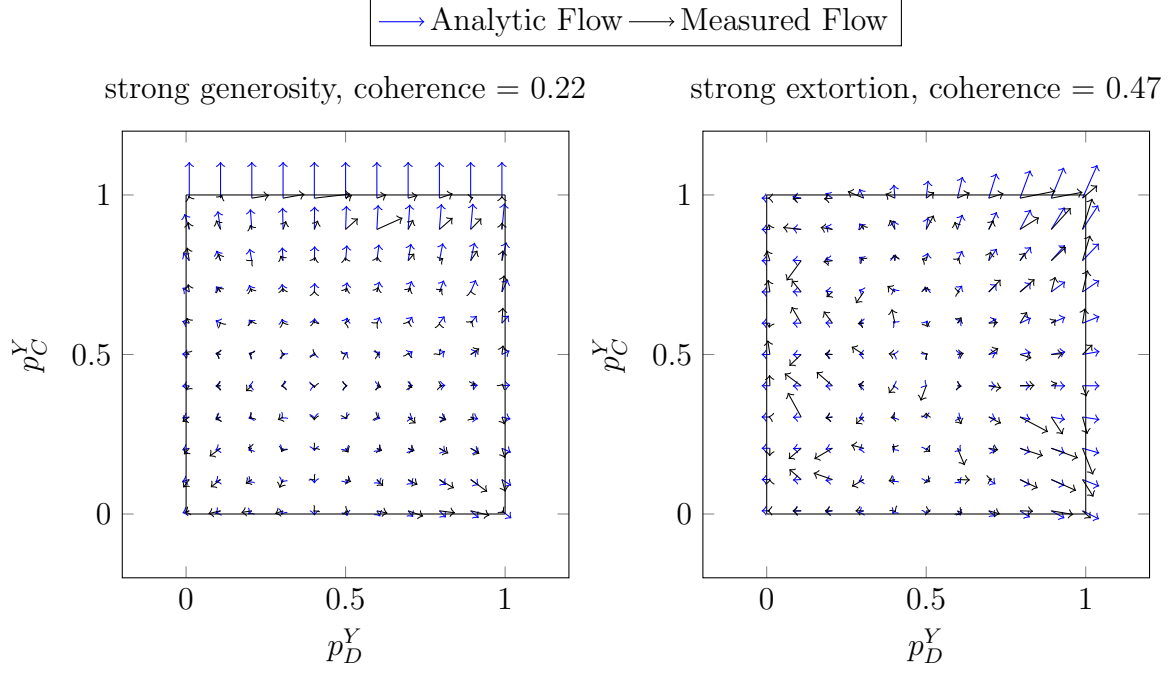
Figure 6: Measurement of the flow for the strong extortion and strong generous strategies with computer generated date sequences of 500 agents and 200 rounds

[3] Henrich J, et al. (2001) In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *Am Econ Rev* 91.

[4] Erev I, Haruvy E (2012) in *The Handbook of Experimental Economics*, eds Kagel JH, Roth AE (Princeton University Press), pp 1–124.

[5] Chammah A, Rapoport AM (1965) *Prisoner's Dilemma - A Study In Conflict And Cooperation* (University of Michigan press), 165 edition.

[6] Rapoport A, Mowshowitz A (1966) Experimental studies of stochastic models for the Prisoner's dilemma. *Syst. Res.* 11:444–458.

[7] Sigmund K (2010) in *Evolutionary Game Dynamics*, ed Sigmund K Vol. 69, pp 1–27.

[8] Helbing D, Lozano S (2010) Phase transitions to cooperation in the prisoner's dilemma. *Phys Rev E Stat Nonlin Soft Matter Phys* 81:2–5.

[9] Axelrod R (1984) *The evolution of cooperation* (Basic Books, New York).

[10] Press WH, Dyson FJ (2012) Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proc Natl Acad Sci USA* 109.

[11] Stewart AJ, Plotkin JB (2012) Extortion and cooperation in the Prisoner's Dilemma. *Proc Natl Acad Sci USA* 109:10134–10135.

[12] Hilbe C, Röhl T, Milinski M (2014) Extortion subdues human players but is finally punished in the prisoner's dilemma. *Nat Commun* 5.

[13] Kraines D, Kraines V (1993) Learning to cooperate with pavlov an adaptive strategy for the iterated prisenor's dilemma with noise. *Theory and Decision* 35:107–150.

[14] Roth A, Erev I (1998) Prediction how people play games: Reinforcement learning in games with unique strategy equilibrium. *Am Econ Rev* 88:848–881.

[15] Fudenberg D, Levine DK (1998) *The theory of learning in games* (MIT Press, Cambridge, London), p 276.

[16] Miller RR, Barnet RC, Grahame NJ (1995) Assessment of the Rescorla-Wagner model. *Psychol Bull* 117:363–386.

[17] Sutton R, Barto A (1990) in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, eds Gabriel M, Moore J (MIT Press, Cambridge, MA, US), pp 497–537.

[18] Bush RR, Mosteller F (1953) A Stochastic Model with Applications to Learning. *Ann Stat* 24:559–585.

[19] Thorndike EL (1898) Animal intelligence: an experimental study of the associative processes in animals. *Psychol Monogr* 2.

[20] Neiman T, Loewenstein Y (2011) Reinforcement learning in professional basketball players. *Nat Commun* 2:569.

[21] Izquierdo SS, Izquierdo LR, Gotts NM (2008) Reinforcement Learning Dynamics in Social Dilemmas. *J Artif Soc Soc Simul* 11.

[22] Xu B, Zhou Y, Lien JW, Zheng J, Wang Z (2015) Extortion outperforms generosity in iterated Prisoners' Dilemma. (`http://arxiv.org/abs/1505.02378`) [Accessed November 10, 2015].

[23] Macy MW, Flache A (2002) Learning dynamics in social dilemmas. *Proc Natl Acad Sci USA* 99 Suppl.:7229–7236.

[24] Hilbe C, Nowak MA, Sigmund K (2013) Evolution of extortion in Iterated Prisoner's Dilemma games. *Proc Natl Acad Sci USA* 110:6913–8.

[25] Rabin M (1993) Incorporating Fairness into Game Theory and Economics. *Am Econ Rev* 83:1181–1302.